Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 60 (2015) 371 – 380

Procedia Computer Science

19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems

Depth Sensor Based Automatic Hand Region Extraction by Using Time-Series Curve and Its Application to Japanese Finger-spelled Sign Language Recognition

Katsufumi Inoue[a,*], Takami Shiraishi[a], Michifumi Yoshioka[a], Hidekazu Yanagimoto[a]

[a]*Graduate School of Engineering, Osaka Prefecture University, 1-1, Gakuencho, Naka, Sakai, Osaka, 599-8531 Japan*

Abstract

Hand sign recognition is one of most challenging issues in computer vision and human computer interaction, and many researchers tackle this issue. In this research, we focus on JFSL (Japanese Finger-spelled Sign Language) which is one of hand signs. The tasks for achieving high performance of JFSL recognition as well as other hand signs are how to extract hand region precisely and how to recognize hand signs accurately. To deal with the former task, in this paper, we propose an automatic hand region extraction method with a depth sensor. The characteristic points of our proposed method are to utilize Time-Series Curve, which is one of contour features, and to extract hand region accurately without wearing landmark object such as a color wristband. On the other hand, to tackle the latter task, in this research, we focus on a deep neural network based recognition method since such a method is reported that it allows us to achieve high performance for various recognition tasks. Therefore, in this paper, we investigate JFSL recognition performance with a deep neural network approach compared to that with the conventional image recognition method (HOG+SVM). From the experimental results with 8 subjects, we have confirmed that our proposed method allows us to extract hand region accurately regardless of subjects and JFSL signs. In addition, from the experimental results with a deep neural network based recognition method for JFSL recognition, we have achieved at least average recognition rate over 88%.

Peer-review under responsibility of KES International

Keywords: Japanese Finger-spelled Sign; Automatic Hand Region Extraction; Time-Series Curve; Depth Sensor; Deep Neural Network

1. Introduction

Recently rapid aging leads to increasing number of people with hearing difficulties and this has become a serious problem. To solve this problem, HSL (hand sign language) recognition has been splotlighted as a key technology to assist the communication with them. Although many researchers have already proposed the HSL recognition methods[1] and they have reported that their methods allow us to achieve high recognition performance, the number of HSL words that they can recognize is restricted in their research. Since HSL recognition is one of most challenging tasks in the some research fields such as computer vision and human computer interaction etc., it is still difficult to

* Corresponding author. Tel.: +81-72-254-7279.
E-mail address: inoue@cs.osakafu-u.ac.jp

Peer-review under responsibility of KES International
doi:10.1016/j.procs.2015.08.145

realize practical recognition system. To solve this problem, in this research, we focus on finger-spelled sign language, which is one of HSL, since the number of signs to master it for communication is much less than that of signs of HSL. From this viewpoint, the purpose of our research is to make a finger-spelled sign language recognition system, especially Japanese finger-spelled sign language recognition system[2], for the first step of smooth communication with the hearing-impaired people.

JFSL (Japanese Finger-spelled Sign Language) is a representation method of Japanese syllabary characters called Hiragana. Although signer needs to represent the characters one by one, JFSL is often utilized for representation of proper nouns. Each sign of JFSL is basically represented as the shape of a gesturing hand and some signs are represented by moving a gesturing hand while keeping the shape of hand for their base sign. The meaning of these signs is changed depending on the movement direction of hand. Since the number of Hiragana is greater than that of alphabet, JFSL recognition is more difficult and challenging task compared with ASL (American Sign Language) recognition. To deal with JFSL, the recognition system is divided into 3 processes; 1) hand region extraction, 2) hand shape recognition, 3) tracking a gesturing hand and recognition of its moving direction. For the first step to realize this system, in this paper, we mainly mention the first process and secondarily mention the second process.

For the first process, recently, depth sensor has been focused on as a key device for hand region extraction[3–5] because the methods with depth sensor enable us to robustly extract hand region for some noises, e.g. illumination change and cluttered background, compared with the color based methods[2,6–8]. However, in order to extract hand region accurately, some of them require signers to represent signs within the specific distance range from the depth sensor and require their hand to be the front-most object from the depth sensor. To relax these restrictions, a method using a depth sensor and a color wristband has also been proposed[4]. However, in such a method, wearing a color wristband on a gesturing hand's wrist is necessary to extract hand region precisely, which is rather inconvenient for real world applications. To solve these problems mentioned above, in this paper, we propose a new depth sensor based hand region extraction method by utilizing TSC (Time-Series Curve)[9], which is one of contour features. The characteristic point of our proposed method is to extract hand region automatically without wearing a landmark item such as color wristband. From the experimental results with 8 subjects for JFSL signs, we have confirmed that our proposed method allows us to extract hand region accurately regardless of subjects and JFSL signs. Additionally, compared with the conventional method with a depth sensor and a black wristband, we have achieved similar extraction results without wearing it on a gesturing hand's wrist. Moreover, in this research, we have investigated JFSL signs recognition accuracy with the hand region images extracted by our proposed method. From the experimental results with a deep neural network based method[10–12], we have achieved at least average recognition rate over 88%.

Although the contributions of our proposed method are little in the research area such as computer vision and human computer interaction, compared with existing work, our proposed method has following contributions.

1. In our proposed method, we employ TSC for extracting hand region precisely with depth sensor. As shown in Fig. 8, by using TSC, we have achieved similar extraction result for hand region without wearing a color wristband compared with the conventional method using a color wristband.
2. To our knowledge, this is the first research of JFSL signs recognition utilizing deep neural network based method. In some of JFSL signs recognition methods, HOG (Histogram of Oriented Gradient)[13] and SVM (Support Vector Machine) are utilized. Compared with these methods, in this research, we utilize simple deep neural network based method, i.e. the network includes convolution, pooling, normalization and activation layers.

The rest of this paper is organized as follows. In Section 2, we introduce related work of hand region extraction and finger-spelled sign language recognition. Next we explain our proposed method in Section 3. In Section 4, we show the experimental results. Finally, we conclude this paper in Section 5.

2. Related Work

Finger-spelled sign language recognition is one of most challenging tasks in computer vision and human computer interaction. In the research of finger-spelled sign language recognition, many researchers mainly deal with AFSL (American Finger-spelled Sign Language)[3,6,7], BFSL (British Finger-spelled Sign Language)[8] or FN (Finger-spelled Number)[4,5,14]. For these work, especially, we handle JFSL (Japanese Finger-spelled Sign Language) because JFSL

recognition task is discussed insufficiently in this research area. To deal with JFSL recognition task, electronic grove based methods [15] and camera based method [16] have already been proposed. Since the former methods require a signer to wear a such expensive device and this is inconvenient for real world application, in this research, we employ the camera based approach. In this approach, there are mainly two tasks; 1) how to extract hand region precisely, 2) how to recognize the finger-spelled signs accurately using the information of the extracted hand region. In this research, we mainly focus on the former task and subsidiarily focus on the latter task. JFSL is represented with the shape of hand region which include from wrist to fingers. In the following, we describe the region including from wrist to fingers as "hand region" and the region including hand and arm as "hand-arm region". In this section, we introduce the conventional hand region extraction method and explain our approach for it. Additionally, we introduce simple finger-spelled sign language recognition approach and recent machine learning approach.

2.1. Related Work of Hand Region Extraction

In the previous camera based work for hand region extraction, methods based on a color grove [6], markers [17], skin color [2,7,8] and depth information [4,5] have already been proposed. In the former two methods, as mentioned above, wearing something such as a grove or markers on a gesturing hand makes the domestic application inconvenient. Additionally, the skin color based methods are not robust for illumination change and background cluttered. Therefore, due to these problems, we utilize depth information for accurate hand region extraction. Although a specific sensor such as depth sensor is necessary for hand region extraction, recently, depth information based methods have been focused on as a hopeful method for this task since a tablet mounted depth sensor such as "Structure Sensor" [1] and "Google Tango" [2] is developed.

As a depth information based method, Ren et al. [4] have proposed a method that extract hand region accurately by detecting the position of wrist with a black wristband after extracting hand-arm region roughly with depth information. Therefore, this method also require signer to wear a color wristband. In addition, in this method, hand is required to be the front-most object from a depth sensor. To relax these problems, our proposed method automatically detects the wrist position of gesturing hand by using TSC [9] without wearing a color wristband. From this process, our proposed method allows us to extract hand region precisely.

2.2. Related Work of Finger-spelled Sign Language Recognition

In the previous work of camera based finger-spelled sign language recognition, the recognition process is divided into 2 steps; feature extraction and matching signs by using extracted features. In the first step, various kind of image features are utilized for feature extraction. For example, in the conventional methods, contour features [6,18], hand skeleton features [19], local features such as HOG (Histogram of Oriented Gradients) [13], etc. are extracted from a hand region image. After these features are extracted, in the second step, matching of represented signs with these features is generally done by using a recognizer like SVM (Support Vector Machines) [6,14] or by using a evaluation metric such as EMD (Earth Mover's Distance) [4,5]. Although these methods have been reported that they can achieve high recognition accuracy, they still have a problem that signs which have similar appearances are difficult to recognize.

For this problem, in this research, we focus on deep neural network approach. After the drastic improvement in image classification task has been reported [11], many researchers utilize the recognizer having such a deep architecture in various recognition tasks, e.g. object classification [20], speech recognition [21], etc. However, in our best knowledge, this is first study of JFSL recognition that employs deep neural network approach. In this research, by simply applying the deep neural network widely used in this research area to JFSL recognition task, we investigate its recognition performance with the hand region images extracted by our proposed method.

[1] http://structure.io/
[2] http://www.google.com/atap/projecttango/#project

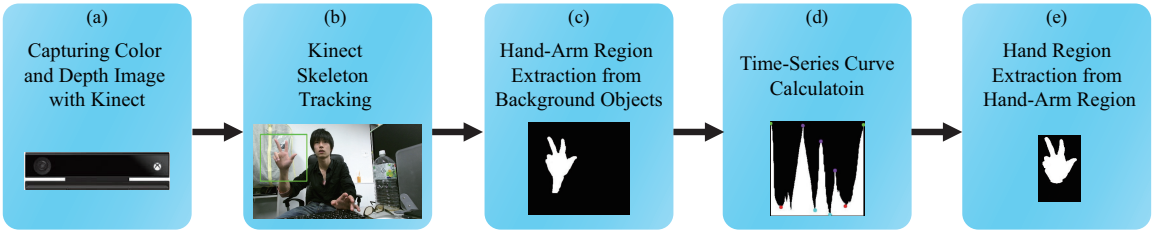
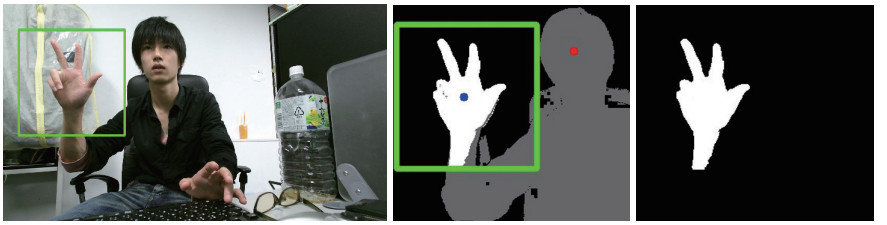


Fig. 1. Overview of hand region extraction process of our proposed method.



(a) (b) (c)

Fig. 2. Extraction of hand-arm region: (a) Detection of approximate hand-arm region as shown green rectangle; (b) Hand-arm region segmentation depending on depth information. Red and blue point show the skeleton position of head and hand calculated with Kinect SDK, respectively; (c) Extraction result of hand-arm region. In this figure, due to the difference of image size of color and depth image, we show (b) and (c) images focused on body and hand regions.

3. Proposed Method

In this section, we explain the detail of our proposed hand region extraction method. As a depth sensor, in this research, we utilize Microsoft Kinect for Windows v2 sensor. In the following, we describe this sensor as "Kinect". Kinect allows us to capture a color image and a depth image simultaneously.

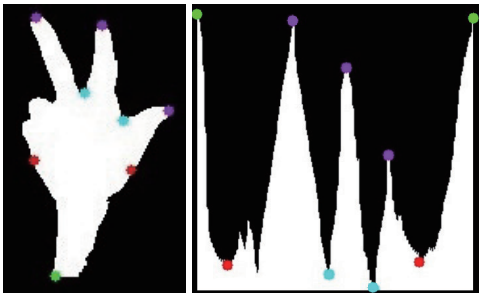
Figure 1 shows the overview of our proposed method. The goal of our proposed method is to obtain a color hand image which includes only hand region, from a color image captured by Kinect. To realize this, first, our proposed method roughly detect the gesturing hand position by using Kinect skeleton tracking which is implemented in Kinect SDK and a hand-arm region is roughly extracted by using depth information of the detected hand position. From this process, we can separate the hand-arm region from background objects. After that, by calculating TSC feature, our proposed method assumes the wrist position without using a color wristband. Finally, we can segment hand region and arm region using information of the assumed wrist position. The concrete process of our proposed method is as follows.

First, in order to separate hand-arm region from background objects, we utilize Kinect skeleton tracking, which allows us to approximately detect the position of hand as shown in Fig 2 (a). In Fig. 2, green rectangle shows the approximate hand-arm region. In this research, the size of green rectangle is set to 30cm $\times$ 30cm, whose center point is the hand position calculated by Kinect skeleton tracking. The green rectangle region still includes hand-arm region and background objects. Therefore we extract hand-arm region from the green rectangle region depending on distance from Kinect, i.e., let D be a distance from Kinect, we extract the region within a distance range Ω from Kinect, which satisfies the following eq.(1).

$$D \leq \frac{D_{\text{hand}} + D_{\text{head}}}{2} \tag{1}$$

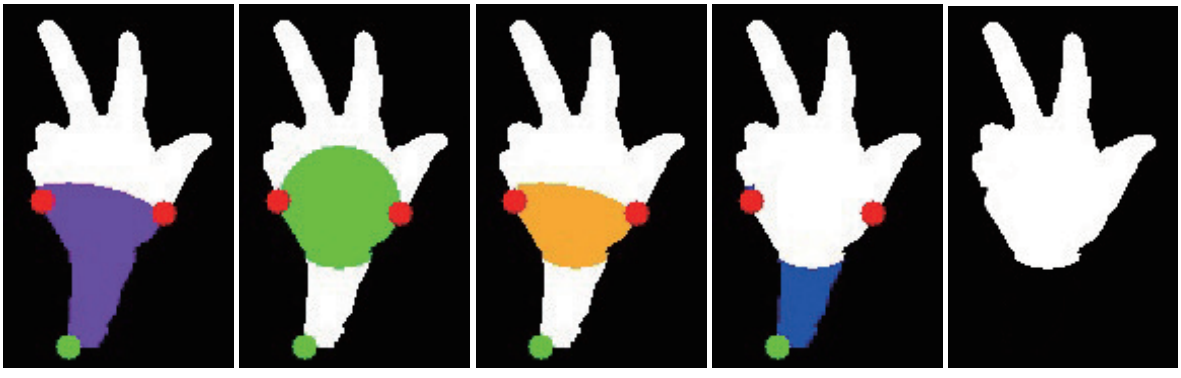
where D_{hand} and D_{head} are distance of hand and head position extracted by Kinect skeleton tracking, respectively. From these processes, we extract hand-arm region from a captured image as shown in Fig. 2 (c).

Next, let us explain the calculation of TSC feature. In our proposed method, we utilize TSC feature in order to detect approximate wrist position from hand-arm region. The detail process of TSC feature calculation is as follows. First, let n_c be the total number of contour point of hand-arm region, we calculate the distance $d_i^c (i \in \{1, 2, \ldots, n_c\})$



(a) (b)

Fig. 3. Example of hand-arm region and its TSC. (a) Example of hand-arm region. Green, red, cyan and purple points show elbow point, wrist points, finger web points and fingertip points, respectively; (b) Its TSC. Each color point corresponds to the point in (a).



(a) (b) (c) (d) (e)

Fig. 4. (a) Region Φ which is near elbow point based on the threshold r_e; (b) Circle region Ψ with a diameter which is the distance between two wrist points; (c) Common region of Φ and Ψ; (d) Difference region Δ of (a) and (c); (e) Result of hand region extraction.

between each contour point and center point of hand-arm region. TSC is the histogram of d_i^c ordered by contour points. Therefore d_i^c for contour points of finger tips and elbow become local maximal value, and d_i^c for contour points of finger webs and wrist is local minimal value. In this research, we define the contour point having the maximum distance D from Kinect in the hand-arm region as "elbow point" as shown green point in Fig 3, and we calculate TSC feature based on this point. Figure 3 (a) and (b) show an example image of a hand-arm region and its histogram as TSC feature, respectively. In this paper, TSC feature as shown in Fig. 3 (b) is normalized based on the maximum and minimum distance d_i^c. From TSC feature, we select 2 contour points having the local minimal d_i^c near elbow point as shown red points in Fig. 3. In the following, we call these points as "wrist points". From this process, we can detect the position of wrist points automatically by using no color wristband.

Finally we explain the segmentation way of hand region from hand-arm region using the information of wrist points. Before we explain the concrete process of hand extraction, let us explain the problem of wrist points detection. As shown in Fig. 3, although we can detect wrist points automatically with the process mentioned above, there is the case that these points are not detected precisely. Therefore, if we simply segment the hand region from hand-arm region based on the wrist points which are detected imprecisely, segmentation performance strongly depend on detection performance of wrist points. In order to relax this problem, we extract hand region from hand-arm region as follows. Let r_e be the longer distance between each wrist point and elbow point. First we segment region Φ as shown purple region in Fig. 4 (a) where the distance from elbow point is within r_e. Additionally, we segment region Ψ as shown green region in Fig. 4 (b) where the points of hand-arm region are included in the circle region whose diameter is the distance between wrist points. Next, from region Φ and Ψ, we obtain region Δ which satisfies eq.(2).

$$\Delta = \Phi - (\Phi \cap \Psi) \tag{2}$$

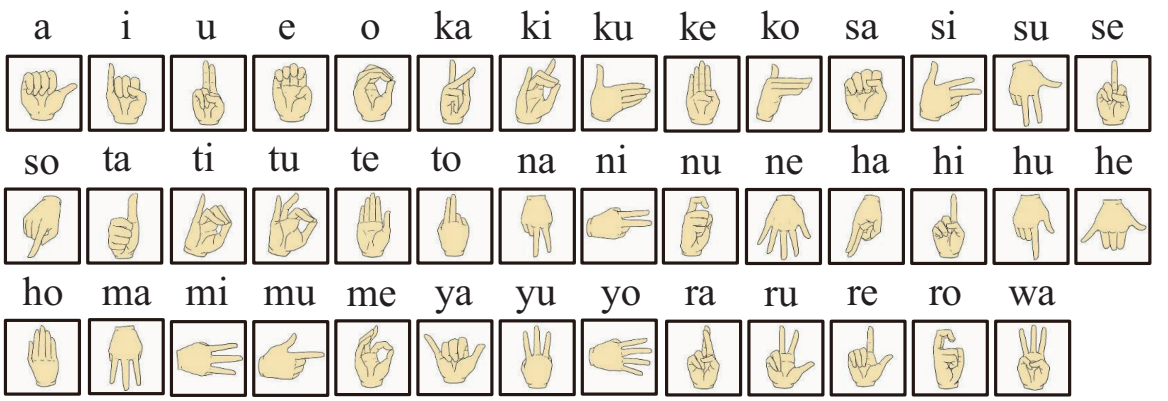


Fig. 5. 41 static JFSL signs.

Figure 4 (c) and (d) show common region of Φ and Ψ and region Δ as yellow region and blue region, respectively. After that, finally, we can obtain the hand region as shown in Fig. 4 (e) by removing region Δ from hand-arm region. From these processes mentioned above, we can robustly extract hand region for position gap between detected wrist points and true wrist points.

4. Experiments

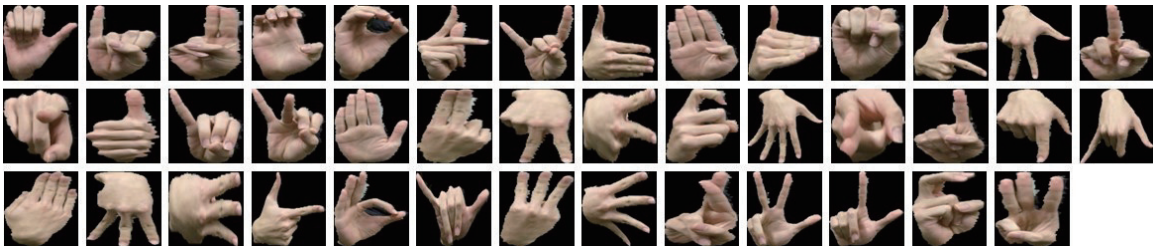
We have evaluated our proposed method with 8 subjects (5 males and 3 females). In this research, we restricted the JFSL signs for evaluation to static signs, i.e. we utilized 41 static JFSL signs which were needed no motion for representation [3] as shown in Fig. 5. Each signs were captured by Kinect 10 times per subject. From this, in the experiments, we utilized 3280 images. For the image size of Kinect, color image size was 1920×1080 and depth image size was 512×424. By using these images, we have evaluated our proposed hand region extraction method. Additionally, we have investigated the performance of JFSL recognition with the deep neural network approach as well as a simple image matching approach (HOG + SVM).

4.1. Experiments of Hand Region Extraction

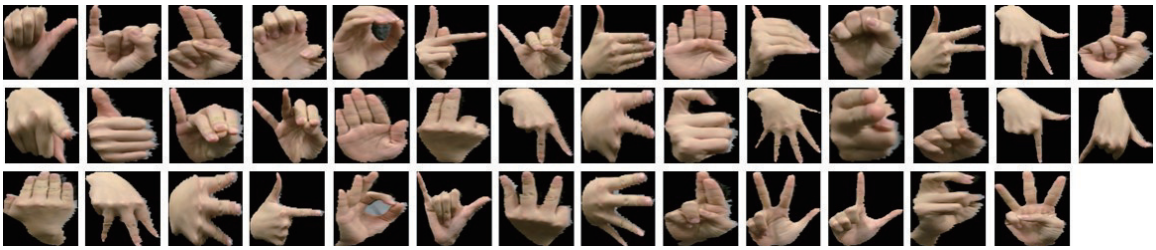
First, we have evaluated the effectiveness of our proposed hand region extraction method. In this experiment, we have investigated the extraction accuracy per sign as well as per subject. Furthermore, we have compared the extraction performance of our proposed method with that of the conventional method which utilizes wearing a black wristband.

Figure 6 shows the hand region extraction results for 41 JFSL signs with subject 4 and 8 who allows us to comparatively achieve the best and worst extraction performance among subjects, respectively. In addition, Fig. 7 shows the result of hand region extraction of each subject in the case of “mu” sign. As shown in these figures, although hand regions of some signs are not precisely segmented from background, we have confirmed that our proposed hand region extraction method enables us to achieve high extraction performance with different subjects as well as different signs. The reason of miss extraction mentioned above is that the recalibration between color and depth information, and the performance of distance calculation of Kinect. For the former reason, in this research, we employed the recalibration method implemented in Kinect SDK. Therefore, in order to solve this problem, we have considered that more accurate recalibration method is required for improvement of extraction performance. On the other hand, the latter problem is caused by the limitation of distance calculation performance for current Kinect. Therefore, we hope that more accurate sensor is developed.

[3] 5 dynamic signs, which are needed a motion for representation, and 41 static signs are included in JFSL signs.



(a) Hand region extraction results with subject 4.



(b) Hand region extraction results with subject 8.

Fig. 6. Hand region extraction results for 41 JFSL signs with subject 4 and 8 who allowed us to comparatively achieve the best and worst extraction performance among subjects, respectively. The size of hand region images is normalized as 256×256.

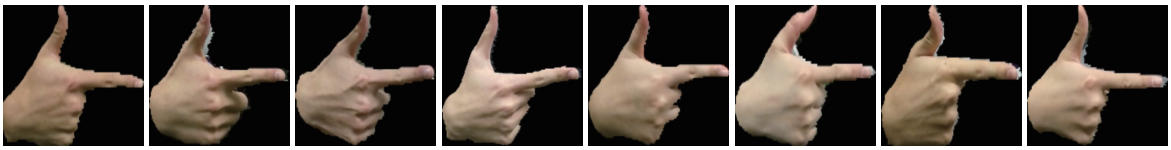
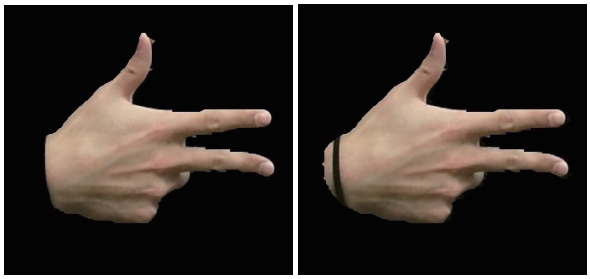


Fig. 7. Result of hand region extraction of each subject in the case of "mu" sign. The size of hand region images is normalized as 256×256.



(a) (b)

Fig. 8. Comparison of hand region extraction result in the case of "shi" sign. (a) and (b) shows the hand region extraction result of conventional and proposed method, respectively.

Table 1. Comparison of processing time for extracting hand region from hand-arm region.

	Proposed Method	Conventional Method
Processing Time [ms]	2.3	1.2

Next, we have compared the proposed method with the conventional method which utilizes wearing a black wrist-band on a wrist of gesturing hand. Figure 8 shows the comparison result of hand region extraction in the case of "shi" sign. Although a part of arm region as well as wrist band region is extracted in our proposed method as shown in Fig. 8,

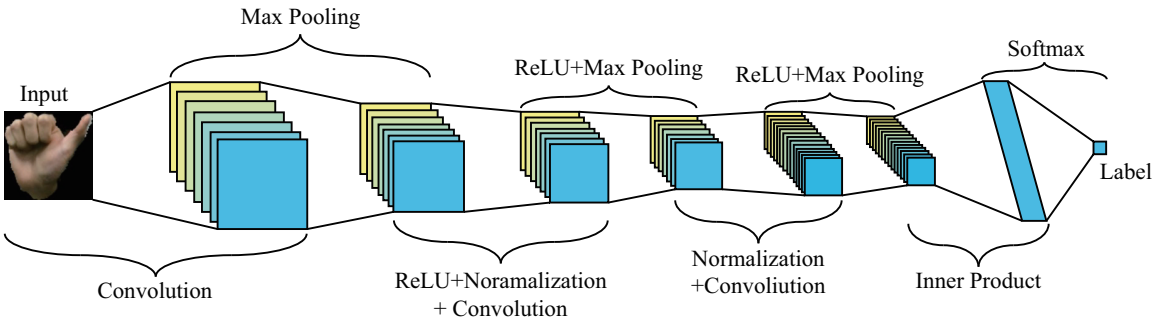


Fig. 9. Network Architecture utilized JFSL recognition in the experiments. We utilized same network that were pre-defined as sample network in caffe library [22] for cifar10 dataset.

the extracted hand region with our proposed method is only 3.4% larger than that with the conventional method. From this result, we have confirmed that our proposed hand region extraction method without wearing a wristband allows us to achieve the similar extraction performance compared with the conventional method with wearing a wristband. Furthermore, Table 1 shows the processing time of hand region extraction for both methods. As shown in Table 1, the difference of extraction time between both methods is approximately 1[ms]. From this result, we have confirmed that our proposed method enables us to extract hand region effectively while keeping the processing speed of hand extraction as fast as possible.

From the results mentioned above, we have confirmed that our proposed method enables us to precisely extract hand region with different signs as well as different subjects by using no landmark item such as a color wristband. In this research, due to the time limitation, quantitative evaluation has not been applied for these experiments. Therefore, this is one of future work.

4.2. Experiments of JFSL Recognition

Next, we have investigated JFSL recognition accuracy by using hand region images extracted by our proposed method. In this experiment, we employed the following two recognition approaches; 1) HOG+SVM 2) deep neural network. In the former approach, a HOG feature was extracted from a single hand region image, whose dimension is 2916. In addition, SVM with linear kernel and 1 vs. 1 approach for learning SVM were utilized for recognition in this experiment. On the other hand, in the latter approach, we utilized the same network structure implemented in Caffe library [22] as sample network for cifar10 classification [4]. Figure 9 shows the concrete network structure. This network includes 3 convolution layers [12], 3 max pooling layers [23,24], 2 local response normalization layers [25], 3 ReLU activation layers [11], 1 inner product layer (a fully connected layer) and 1 softmax loss layer. In this experiment, the hand region images for 7 subjects and those for 1 subject were utilized as training data and test data, respectively. By changing the data of test subject, we have investigate the JFSL recognition accuracy. Additionally, the size of hand region images is normalized as 256×256 and 64×64. For experiments with HOG+SVM, we employed both image size. On the other hand, for experiments with deep neural network, only the images whose size is 64×64 were utilized due to the memory limitation of our GPU processing unit. In following of this paper, we describe the method with HOG+SVM approach as "H+S" and that with deep neural network as "DNN".

Figure 10 shows the experimental results where the average recognition rate for both methods. From Fig. 10, even though long processing time is necessary for training network in DNN, we have confirmed that DNN allows us to achieve higher recognition accuracy for JFSL recognition. In addition, as shown in Fig. 10, we have considered that the difference of recognition accuracy between DNN(150k) and DNN(Best) is comparatively large. Possible reasons of this problem are that the training process iteration for the network and the investigation of parameter for network training are insufficient. Moreover, in this experiment, although we simply utilized pre-defined network

[4] http://www.cs.toronto.edu/~kriz/cifar.html

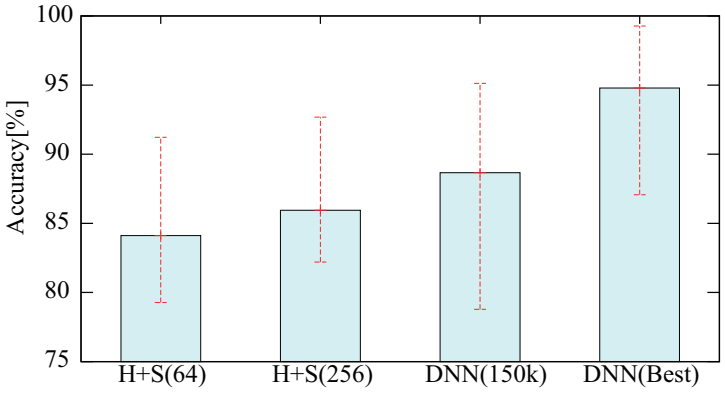


Fig. 10. Average JFSL recognition accuracy with H+S(HOG + SVM) and DNN(Deep Neural Network) approaches. In the 2 left results, 64 and 256 means the size of images utilized for experiments. On the other hand, in the 2 right results, DNN(150k) and DNN(Best) means that the recognition results in the case of using the network constructed after 150k times learning process iterated and the best recognition result in the case of using the network constructed during 150k times learning process, respectively. The top and bottom error bar shows the best and worst recognition performance among 8 subjects, respectively.

in Caffe library, we have no knowledge whether the structure of this network is optimal for achieving higher JFSL recognition performance. Accordingly, from these results, we have considered that more investigation of network training conditions as well as network structure is necessary for improving the recognition accuracy of JFSL signs in the future work. From these results, although many tasks we must tackle are remained, we have confirmed that H+S and DNN enables us to recognize JFSL signs with at least 84% and 88% of average rate by using the hand region images extracted our proposed method, respectively. Therefore we have considered that our hand region extraction method is effective for JFSL recognition.

5. Conclusion

In this paper, we proposed automatic hand region extraction method by using TSC and we investigated the recognition performance for JFSL recognition with DNN approach by using the hand region images extracted by our proposed method. From the experimental results, we have confirmed that our hand region extraction method allows us to achieve high extraction performance with different subjects as well as different signs of JFSL without wearing a wristband on a gesturing hand. Additionally, we have confirmed that our proposed method enables us to achieve similar extraction performance compared with the conventional method which utilizes wearing a black wristband on a gesturing hand. From the experimental results of 41 JFSL signs recognition, we have achieved at least average recognition rate over 88% and best recognition rate 99% with DNN approach.

Future work is to improve the recalibration between color and depth information, to evaluate our proposed hand region extraction method quantitatively. Additionally, to investigate more effective network training condition for JFSL recognition in DNN is also one of our future work.

References

1. Mitra, S., Acharya, T.. Gesture Recognition: A Survey. *IEEE Transactions on Systems Man and Cybermetics (C) Applications and Reviews* 2007;**37**(3):311–324.
2. Terrillon, J.C., Pilpre, A., Niwa, Y., Yamamoto, K.. Robust Face Detection and Japanese Sign Language Hand Posture Recognition for Human-Computer Interaction. In: *Proc. of ICVI2002*. 2002, p. 369–376.
3. Kuznetsova, A., Leal-Taixé, L., Rosenhahn, B.. Real-Time Sign Language Recognition Using a Consumer Depth Camera. In: *Proc. of ICCVW2013*. 2013, p. 83–90.
4. Zhon Ren Junsong Yuan, J.M., Zhang, Z.. Robust Part-Based Hand Gesture Recognition Using Kinect Sensor. *IEEE Transactions on Multimedia* 2013;**15**(5):1110–1120.
5. Chong Wang, Z.L., Chan, S.C.. Superpixel-based Hand Gesture Recognition with Kinect Depth Camera. *IEEE Transactions on Multimedia* 2015;**17**(1):29–39.

6. Daniel Kelly, J.M., Markham, C.. A Person Independent System for Recognition of Hand Posture Used in Sign Language. *Pattern Recognition Letters* 2010;**31**:1359–1368.
7. Pugeault, N., Bowden, R.. Spelling It Out: Real-Time ASL Fingerspelling Recognition. In: *Proc. of ICCV2011 Workshop*. 2011, p. 1114–1119.
8. Liwicki, S., Everingham, M.. Automatic Recognition of Fingerspelled Words in British Sign Language. In: *Proc. of CVPR for HCBA2009*. 2009, p. 50–57.
9. Keogh, E., Wei, L., Xi, X., hee Lee, S., Vlachos, M.. LB_Keogh Supports Exact Indexing of Shapes under Rotation Invariance with Arbitrary Representations and Distance Measures. In: *Proc. of VLDB, 2006*. 2006, p. 882–893.
10. Le, Q.V., Ranzato, M., Monga, R., Devin, M., Chen, K., Corrado, G.S., et al. Building High-level Featurs Using Large Scale Unsupervised Learning. In: *Proc. of ICML2012*. 2012, p. 81–88.
11. Krizhevsky, A., Sutskever, I., Hinton, G.E.. ImageNet Classification with Deep Convolutional Neural Networks. In: *Proc. of NIPS2012*. 2012, p. 1097–1105.
12. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.. Gradient-Based Learning Applied to Document Recognition. *Proc of IEEE* 1998; :2278–2324.
13. Dalal, N., Triggs, B.. Histograms of Oriented Gradients for Human Detection. In: *Proc. of CVPR2005*. 2005, p. 886–893.
14. Dardas, N.H., Georganas, N.D.. Real-Time Hand Gesture Detection and Recognition Using Bag-of-Features and Support Vector Machine Techniques. *IEEE Transactions ofn Instrumentation and Measurement* 2011;**60**(11):3592–3607.
15. Laura Dipietro, A.M.S., Dario, P.. A Survey of Glove-Based Systems and Their Applications. *IEEE Transactions on Systems, Man and Cybermetics (C) Applications and Reviews* 2008;**38**(4):461–482.
16. Yamashita, T., Watasue, T.. Hand Posture Recognition Based on Bottom-up Structured Deep Convolutional Neural Network with Curriculum Learning. In: *Proc. of ICIP2014*. 2014, p. 853–857.
17. Zhao, W., Chai, J., Xu, Y.Q.. Combining Marker-based Mocap and RGB-D Camera for Acquiring High-fidelity Hand Motion Data. In: *Proc. of ACM SIGGRAPH/ESCA2012*. 2012, p. 33–42.
18. Belongie, S., Malik, J., Puzicha, J.. Shape Matching and Object Recognition Using Shape Contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2002;**24**(4):509–522.
19. Bai, X., Latecki, L.J.. Path Similarity Skeleton Graph Matching. *IEEE Transactions on Pattern Analysis and Machine Ingelligence* 2008; **30**(7):1282–1292.
20. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., et al. Going Deeper with Convolutions. *CoRR* 2014;**abs/1409.4842**.
21. Hinton, G., Deng, L., Yu, D., rahman Mohamed, A., Jaitly, N., Senior, A., et al. Deep Neural Networks for Acoustic Modeling in Speech Recognition. *IEEE Signal Processing Magazine* 2012;**29**(6):82–97.
22. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., et al. Caffe: Convolutional Architecture for Fast Feature Embedding. *arXiv preprint arXiv:14085093* 2014;.
23. Serre, T., Wolf, L., Poggio, T.. Object Recognition with Features Inspired by Visual Cortex. In: *Proc. of CVPR2005*. 2005, p. 994–1000.
24. Banzato, M., Boureau, Y.L., LuCun, Y.. Sparse Feature Learning for Deep Belief Networks. In: *Proc. of NIPS2007*. 2007, p. 1185–1192.
25. Pinto, N., Cox, D.D., DiCarlo, J.J.. Why is Real-World Visual Object Recognition Hard? *PLoS Computational Biology* 2008;:151–156.